IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 5, 2020, accepted August 21, 2020, date of publication August 26, 2020, date of current version September 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3019537

Optimal Data Collection for Mobile Crowdsensing Over Integrated Cellular and Opportunistic Networks

DOAA MOHSIN MAJEED, LIN ZHANG, AND KE SHI, (Member, IEEE)
School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

Corresponding author: Ke Shi (keshi@hust.edu.cn)

This work was supported in part by the China Education and Research Network (CERNET) Innovation Project under Grant NGII20170323, and in part by the Natural Science Foundation of Hubei Province under Grant 2016CFC721.

ABSTRACT One of the main challenges that mobile crowdsensing systems must solve is reducing data collection costs while still holding high data delivery probability. Compared with cellular networks, opportunistic networks can significantly reduce data transfer costs at the cost of damaging data delivery probability. This paper proposes an optimal data collection scheme for mobile crowdsensing, which utilizes integrated cellular and opportunistic networks to implement data collection. We use data collecting path to describe how the sensing data are collected and sent to the back-end platform, though cellular networks directly or through multi-hop opportunistic networks. An optimal data collection problem is then formulated as choosing specific data collecting paths from candidate path set to minimize the total crowdsensing cost under the data delivery constraints, which can be considered as a minimum set covering problem. To solve this NP-hard problem, we design and implement a greedy heuristic algorithm that constructs the solution in multiple steps by making a locally optimal decision in each step. We conduct extensive simulations based on three real-world traces: Cambridge, Infocom06, and UPB. The results show that, compared with other data collection approaches, our approach achieves a better tradeoff between cost and data delivery.

INDEX TERMS Data collection, mobile crowdsensing, opportunistic networks, cellular networks.

I. INTRODUCTION

Nowadays, mobile crowdsensing become more and more popular with the development of mobile personal devices such as smartphones or smartwatches with significantly more sensing, computing, communication, and storage resources [1], [2]. With the help of these devices, data related to the environment, transportation, healthcare, safety, and so on can be collected without deploying sensors in-situ. Every mobile device user can be the potential participant of sensing activities, and the inherent mobility of participants provides unprecedented sensing area coverage. Mobile crowdsensing is a technology that allows large scale, cost-effective sensing of the physical world. Now many mobile crowdsensing applications such as environmental quality monitoring [3], noise pollution assessment [4] and traffic monitoring [5], [6] have been developed to collect data from the field, transfer the collected data using commonly available communication technologies to back-end platform, typically located in the cloud, for data processing and analysis, and provide service to the users.

An essential operation in mobile crowdsensing is to perform data collection to minimize the communication cost while still satisfying sensing coverage and data quality constraints. Designing such a data collection strategy usually includes (1) deciding how to recruit the participants among the candidates who visit the sensing area and have the willingness to capture data; (2) deciding how to transfer the captured data to the back-end platform.

By now, lots of research efforts [7]–[11] have been conducted in designing efficient participant recruitment strategies. In these researches, participant recruitment is formatted as an optimization problem of maximizing or minimizing a real objective function by systematically choosing input values from within an allowed set and computing the objective function's value. For example, the optimization problem can be minimizing the total cost under a predefined coverage constraint [8], selecting a predefined number of participants to maximize the spatial coverage [9], [10], or maximizing spatial coverage under total cost constraint [11].

The associate editor coordinating the review of this manuscript and approving it for publication was Martin Reisslein.

However, most of these existing methods adopt a very straightforward data transferring strategy. The participants transfer data to the back-end platform using the cellular network as soon as their devices' sensors generate data. The cellular network operators provide more and more cheap data plan options with the development of 3G/4G. However, the participants may still need to pay extra money for data transferring, which may prevent mobile users from participating in sensing activities and lead to more incentive costs. This strategy generates additional workload for the cellular network and increases the communication cost. To reduce the cost, piggyback based approaches [8]–[12] are proposed to leverage the opportunities for collecting and transferring sensor data that frequently occur during everyday smartphone user operations, such as placing calls or using applications. Hierarchical structure [29] can be constructed to reduce the interference caused by extra network traffic and low energy consumption. Data aggregation [27] and reduction technologies such as compression and spatial-temporal fusion [5] are also utilized to reduce the amount of data needed to be transferred. These methods can only solve part of the problem. When sensing activities generate large amounts of data such as photos, audios, or videos, the participating cost for an individual user may still be high.

With the development of short-distance wireless communication technologies [13], [14], mobile devices can form opportunistic networks to communicate without using cellular networks. Sensing data are collected from mobile devices and transmitted to the back-end platform through opportunistic networks, especially for the large volume data for which it is expensive. Many crowdsensing systems adopt opportunistic networks rather than cellular networks to reduce communication costs [15]–[17]. However, data collection through opportunistic networks still has many challenges. One of the main challenges is on successful data delivery probability. Compared with cellular networks, the connections between mobile devices and back-end platform depend on intermittent contacts, making it difficult to achieve a high data delivery probability. Although data replication and data redundancy [13] can improve the data delivery probability, it increases the amount of transmitted data and leads to high resource consumption.

Unlike solely relying on cellular networks or opportunistic networks to transfer data to the back-end platform, we propose utilizing integrated cellular and opportunistic networks to implement optimal data collection in mobile crowdsensing. Specifically, mobile devices participating in sensing activities can either transmit data to the back-end platform directly through cellular networks or transfer data to other devices and then let other devices transmit the data to the back-end platform through short-distance radio to balance the cost and successful data delivery probability. To avoid the disadvantages of integrating two networks such as high delivery cost caused by cellular networks and low delivery probability caused by opportunistic networks, we define the data delivery cost and probability model and set the optimizing goal as minimizing the total cost under the data delivery constraints.

The main contributions of this paper are as follows:

(1) We use data collecting paths to define how the sensing data are collected and sent to the back-end platform over integrated cellular and opportunistic networks. The method for calculating the costs and data delivery probabilities of these paths are also given.

(2) We formulate optimal data collection as a minimum set covering problem. Specific data collecting paths are chosen from the candidate path set to minimize the total crowdsensing cost under the data delivery constraints.

(3) The minimum set covering problem is a well-known NP-hard problem. A greedy heuristic algorithm is proposed to get an approximate optimal solution in multiple steps by making a locally optimal decision in each step.

(4) We conduct extensive simulations based on three real-world traces. The results show that the proposed scheme achieves a better tradeoff between delivery cost and probability.

The rest of this paper is organized as follows. Section II reviews related work and Section III gives the system model and problem definition. Finding data collection paths and constructing candidate paths set are presented in Section IV. A heuristic algorithm is given in Section V. Section VI evaluates the performance of the proposed algorithm, and Section VII concludes the paper.

II. RELATED WORK

Data collection is an essential part of building a mobile crowdsensing system [18], which includes the following two aspects: (1) which users should be recruited to participate in sensing activities and (2) how to transfer the sensing data to the back-end platform.

Some research treats user recruitment as some kinds of optimization problems with particular optimized objectives. Reference [19] aims to maximize the coverage under the cost constraints and proposes an approximation algorithm to solve it. Reference [8] aims at minimizing the overall cost with guaranteed spatial and temporal coverage. Reference [39] proposes a participant service quality-aware data collecting mechanism with high coverage, aiming at maximizing service quality and data coverage under the condition of a limited platform budget.

Reference [41] investigates the user recruitment problem in sparse MCS, which can recruit a small number of users to sense data from only a few subareas, and, then, infers the data of un-sensed subareas. It studies the user recruitment problem on both user and subarea sides and proposes a three-step user recruitment strategy. Reference [42] develops a user recruitment system for efficient photo collecting in mobile crowdsensing, where a user recruitment strategy is devised to recruit the optimal k users for finishing the sensing task.

Deep reinforcement learning-based approaches [38], [40] are also proposed to solve the user recruitment problems to improve energy efficiency, data collection ratio, and geographic fairness.

All these methods try to balance the cost and service quality. However, they all use very elementary data transferring strategy; sensing data is transmitted to the back-end platform directly using cellular networks.

To reduce the cost caused by cellular transmission, Piggyback CrowdSensing (PCS) [12] predicts mobile devices usage activities (which is called as Smartphone App Opportunities) such as placing phone calls or browsing the web and exploits these activities to obtain and upload sensing data. An architecture and corresponding algorithms are designed to maximize the benefit possible from smartphone app opportunities. CrowdRecruiter [9] is a user recruitment framework operating on top of PCS, minimizing incentive payments/cost by selecting a small number of participants while still satisfying probabilistic coverage constraints. It first predicts each mobile user's call and coverage probability, then proposes a utility function to measure the joint coverage probability of multiple users, and finally deploys a low-complexity but effective algorithm to select the participants incrementally. CrowdTasker [11] also operates on top of PCS, aiming to maximize the sensing task's coverage while satisfying the cost constraint. However, PCS still uses cellular networks to forward data.

With the development of opportunistic networking, integrating opportunistic networking mechanisms in cellular networks can reduce significantly resource consumption. For example, data offloading can reduce cellular downlink traffics [26]. A novel transmission scheduling method [30] is also presented to reduce the uplink resource consumption by selecting single-hop traditional, opportunistic cellular and opportunistic D2D-aided cellular mode for each data fragment. It is similar to our idea to select a data collecting path for the data sensed by the devices.

Some research focuses on realizing crowdsensing over opportunistic networks that have been successfully used in many applications [28]. Their goal is to reduce the data uploading cost. Due to the mobile users' mobility, which leads to the intermittent link connectivity, sensing data uploading is analogous to the opportunistic network data routing. Reference [16] proposes a participant recruitment and data collection framework operating in Delay Tolerant Network (DTN) mode. The feasibility of several DTN data routing approaches, including epidemic routing, PROPHET, spray and wait, profile-cast, and opportunistic geocast, are investigated, and comprehensive analysis of their performance is provided. Reference [20] proposes an Accept aNd Tolerate (ANT) routing protocol to implement data collection in a social environment with selfish individuals. Besides the devices' contact caused by mobility, it also considers the devices' willingness to cooperate in devices selection. Reference [21] proposes a cooperative data collection framework, where data collectors cooperate with mobile users to send data back to requesters. However, these methods only solve the problem of how to send data to the back-end platform. In mobile sensing, user recruitment must be considered at the same time to achieve optimal data collection.

Reference [7], [15], and [17] address the problem of user selection and data transmission based on the opportunistic networking paradigm at the same time. It aims to collect location-based data, while users, depending on their mobility patterns, may undertake different roles, i.e., sense or relay the original sensor data. Reference [17] formulates the problem under both deterministic and stochastic user mobility as instances of the minimum cost set cover problem with submodular objective functions, designs practical greedy heuristics to solve the problem, and derive the approximation ratios they achieve. The probabilities of opportunistic uploading paths are determined by using user' mobility patterns in the past. Reference [7] proposes two trajectories predicting models, deterministic and probabilistic model. Based on these models, optimal user recruitment is formulated as a linear programming problem aiming to minimize the overall recruitment cost.

Reference [22] and [23] focus on a self-organized mobile crowdsensing. Sensing data is sent from the participants to the individual data requesters directly instead of uploading them to a back-end platform. Data requesters publish a sensing task to sense the specific data for an area. The users entering the area could be recruited, take the sensing data, and forward the data to the requester. UROC (User Recruitment strategy for self-Organized mobile Crowdsensing) [23] estimates the expected profit of recruiting a user, compares the profit with the recruiting cost, and decides whether to recruit the user.

PURE (Prediction-based User Recruitment for mobile crowdsEnsing) [22] divides the users into two groups according to different costs: Pay as you go (PAYG) and Pay monthly (PAYM). PAYM users use cellular networks to upload data, and PAYG users forward the data to PAYM users. It uses a semi-Markov model to determine the probability distribution of user arrival time at a specific area and get the inter-user contact probability. First PAYM users with the largest contact probability entering the sensing area are recruited. Then some PAYG users with the higher contact probability entering the sensing area and higher contact probabilities with PAYM users are recruited. The optimized objective is minimizing cost.

A bio-inspired data transfer framework, bioMCS, deployed over a fog computing platform, is proposed to enforce collaborative crowdsensing among proximate users [37]. It constructs a biological network called transcriptional regulatory network and restricts device energy overhead by taking advantage of energy-efficient D2D communications like WiFi direct data transfer via group owner.

Unlike the above research works, we focus on optimal data collection over integrated cellular and opportunistic networks. Participants send data to a back-end platform that provides data services to data requestors, providing a global view of the monitored areas and supporting comprehensive data analysis. Unlike most existing methods based on one data uploading mode, cellular networking mode, or opportunistic networking mode, the participants use mixed mode to upload the data. PURE divides the users into two groups,

one group uses cellular networking mode, and the other group uses opportunistic mode. Our method does not have such a pre-fixed division. Participant selection and corresponding data uploading mode are determined by optimized problem-solving.

III. SYSTEM MODEL

In this section, we describe the scenario we focus on and present the system model. We also introduce the definition of a data collecting path.

A. SCENARIO AND MODEL DESCRIPTION

The two main actors in the model we consider are the mobile users and a back-end platform running on the cloud that organizes the mobile crowdsensing campaign. A crowd of mobile users, denoted by $U = \{u_1, u_2, u_3, \ldots, u_N\}$, moves around in the sensing area S. There are some Points of Interest (PoIs) denoted by $L = (l_1, l_2, l_3, \ldots, l_M)$ within the sensing area. When a mobile user visits a PoI, it can collect data and upload it to the back-end platform. The back-end platform stores and processes the collected data and provides data services to the subscribed users. To avoid data out of date, we define data collecting cycle T_c. In each T_c, each PoI's data needs to be acquired and uploaded to the back-end platform by at least one mobile user.

The mobile user has two options to upload data. The sensing area is fully covered by a set of cellular base stations denoted by *BSS* (*e.g.*, 3G NodeB or 4G LTE eNB). The first option is sending data to the back-end platform directly through these cellular base stations. There are also a set of wireless access points $WA = (wa_1, wa_2, wa_3, \ldots, wa_K)$ scattered across the area and accessible to the mobile users for uploading data. The mobile users may also form opportunistic networks through D2D communication to extend the *WA*'s coverage. The second option is sending data to the back-end platform through opportunistic networks connecting to the access points belonging to the *WA*.

In mobile crowdsensing, participating devices need to consume computing, storing, communicating, and sensing resources to perform the tasks. Notably, the energy the devices spend in sensing and uploading data may drain their battery faster, and extra pay may be caused if they upload the sensing data to the back-end system through cellular networks. Each user participates in a particular crowdsensing task at a particular cost. Here the cost is mainly determined by the energy consumption and data transferring price.

B. DATA COLLECTING PATH

We use data collecting path to define how the data of PoIs are collected and sent to the back-end platform. There are two kinds of data collecting paths: cellular and opportunistic. A cellular path can be defined as $p_i^{lj} = \{l_j\colon u_k\colon bss_e\}$, where l_j represents point of interest j, u_k represents user k, bss_e represents a specific cellular base station. It means user k can visit PoI j, get its data, and use the cellular networks to send data to the back-end platform. There is only one user along the cellular path. This user is responsible for both data acquisition and uploading. Opportunistic path defines how the data from a PoI reach a *wa* through an opportunistic network realized through the users' mobility and their time-ordered encounters. An opportunistic path can be represented as $p_i^{lj} = \{l_j\colon u_k, \ldots, u_o\colon wa_e\}$, where l_j represents point of interest j, $u_k, \ldots, u_o$ represent a subset of users forming the opportunistic network, wa_e represents an access point.

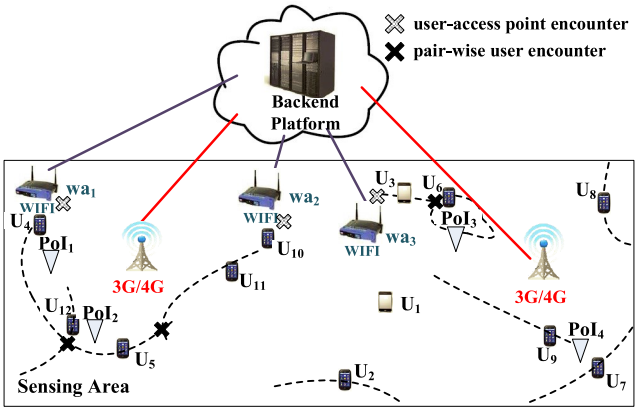

FIGURE 1. Example of a crowdsensing system over integrated cellular and opportunistic networks.

Fig. 1 shows an example of a crowdsensing system over integrated cellular and opportunistic networks. 12 users are roaming around the sensing area covered by cellular networks. Besides cellular networks, users can also upload data to the back-end platform through 4 WIFI access points. The data of 4 PoIs need to be collected. The users visiting the PoIs could be candidate participants. For example, user 5 can get the data of PoI 2 and has three uploading options which correspond to one cellular path $\{l_2\colon u_5\colon bss_e\}$ and two opportunistic paths, $\{l_2\colon u_5\colon wa_1\}$ and $\{l_2\colon u_5, u_{11}, u_{10}\colon wa_2\}$.

The cost and data delivery probability of a data collecting path is denoted by c_p and q_{pl}. For a cellular path, only one mobile user is involved, the cost is the sum of data sampling cost and data transferring cost through the cellular networks:

$$c_p = c_s + c_e. \tag{1}$$

The data delivery probability is

$$q_{pl} = g_{Tc}(u, l)q_c, \tag{2}$$

where $g_{Tc}(u, l)$ is the probability of user visiting PoI in the data collecting cycle, and q_c is the successful transmission probability of the cellular networks. We assume the cellular networks are reliable and q_c is close to 1.

For an opportunistic path, the cost is calculated as

$$c_p = \sum\nolimits_{u \in p} c_u + c_s, \tag{3}$$

where c_s is the data sampling cost, and c_u is the data transferring cost of each user constructing the opportunistic path. The data delivery probability is

$$q_{pl} = g_{Tc}(u, l)q_{wa}, \tag{4}$$

where $g_{Tc}(u, l)$ is the probability of user visiting POI in the data collecting cycle, and q_{wa} is the probability of finding an opportunistic path and successfully transferring data along this path after visiting this POI in this data collecting cycle. The detail of calculating this probability will be given in the next section.

In general, sampling data mainly consumes devices' energy and memory, leading to low cost. Data transferring through cellular or opportunistic networks cause higher energy consumption than sampling. Therefore c_u and c_e are higher than c_s. Cellular network transmission may be charged a fee. So c_u is less than c_e. However, the data delivery probability of opportunistic path is always less than the data delivery probability of cellular path.

C. PROBLEM FORMULATION

Optimized data collection can be described as selecting specific data collecting paths to minimize the cost while satisfying the PoIs' data delivery constraints, as shown in Formula 5. For a PoI, data delivery constraint means the data can be obtained and uploaded to the back-end platform with a certain probability higher than an application determined threshold. The formula details will be given in Section V.

$$\begin{aligned} &min \sum\nolimits_{p\in P} y_p c_p \\ &s.t.\ y_p \in \{0,1\}, \quad \forall p \in P \\ &\quad D_{l,DCP_l} \geq D_{thre}, \quad \forall l \in L \end{aligned} \tag{5}$$

In the example shown in Fig.1, two cellular paths, $\{l_2: u_5: bss_e\}$ and $\{l_2: u_{12}: bss_e\}$, and two opportunistic paths, $\{l_2: u_{12}: wa_1\}$ and $\{l_2: u_5, u_{11}, u_{10}: wa_2\}$, can be used to get the data of PoI 2. If the required data delivery probability is high, the cellular path may be selected to collect data at a high cost. Otherwise, an opportunistic path may be selected to reduce the cost. It is also possible to select two or more paths to get the data of one PoI. For example, the data delivery probabilities of paths $\{l_2: u_5: wa_1\}$ and $\{l_2: u_5, u_{11}, u_{10}: wa_2\}$ are 0.5 and 0.3 respectively. If we assume the users' mobility is independent, the data delivery probabilities of these two paths are independent since users' intersection is empty. The final delivery probability is 0.65 (1-(1-0.5)(1-0.3)) when these two paths are both selected to collect the data of PoI 2.

In our model, data collecting paths are either cellular or opportunistic. Hybrid data collecting paths that use opportunistic networks ending up in a cellular link to a base station are not allowed. We assume all the devices have the same cellular transferring cost c_e, and all the sensed data should be uploaded to the back-end system. The cost of a hybrid path will be the sum of cellular transferring cost and opportunistic transferring cost. Using a hybrid path may reduce the individual cost for a specific device since it does not use the cellular network. Nevertheless, the last device along this path will use the cellular networks to transfer the data to the back-end system. The total cost will be higher than using a cellular path only. If different devices have different cellular transferrin costs, using hybrid data collecting paths may lower the total cost. We plan to consider heterogeneous cost model and hybrid data collecting paths in the future work.

TABLE 1. Encounter probabilities in a time interval.

	u_1	u_2	...	u_N	
u_1	$g_i(u_1,u_1)$	$g_i(u_2,u_1)$	...	$g_i(u_N,u_1)$	$G_i(U,U)$
u_2	$g_i(u_1,u_1)$	$g_i(u_2,u_2)$	...	$g_i(u_N,u_2)$	
...	...	...	...	...	
u_N	$g_i(u_1,u_N)$	$g_i(u_2,u_N)$	...	$g_i(u_N,u_N)$	
l_1	$g_i(u_1,l_1)$	$g_i(u_2,l_1)$	...	$g_i(u_N,l_1)$	$G_i(U,L)$
l_2	$g_i(u_1,l_2)$	$g_i(u_2,l_2)$	...	$g_i(u_N,l_2)$	
...	...	...	...	...	
l_M	$g_i(u_1,l_M)$	$g_i(u_2,l_M)$	...	$g_i(u_N,l_M)$	
w_1	$g_i(u_1,w_1)$	$g_i(u_2,w_1)$	...	$g_i(u_N,w_1)$	$G_i(U,W)$
w_2	$g_i(u_1,w_2)$	$g_i(u_2,w_2)$	...	$g_i(u_N,w_2)$	
...	...	...	...	...	
w_K	$g_i(u_1,w_K)$	$g_i(u_2,w_K)$	...	$g_i(u_N,w_K)$	

IV. CONSTRUCTING CANDIDATE PATHS SET

This section describes how to construct candidate data collecting paths set and calculate the data delivery probability of the corresponding path contained in the candidate paths set.

We divide the data collecting cycle into T_n time intervals. In each time interval t_i, the encounter probabilities of user-user, user-PoI, and user-WA denoted by $g_i(u, u)$, $g_i(u, l)$, and $g_i(u, w)$ respectively can be derived by analyzing users' historical trace. When we cannot get users' phone call traces (including user id, call time, and cell tower) or GPS-coordinates of each user's mobile route from wireless service providers due to the concerns like privacy, we utilize the mobility pattern discovered by the previous research [35], [36] to determine encounter probabilities. We model the contact process between devices as an independent Poisson process and the contact duration between devices as a Pareto distribution. The density and distribution of users in a speicific area can be utilized to determine the model's parameters and derive the encounter probabilities of user-user, user-PoI, and user-WA.

As shown in Table 1, these values can be organized into three matrices including a $N \times N$ matrix $G_i(U, U)$ representing users' encounters, a $N \times M$ $G_i(U, L)$ matrix representing users visiting PoIs, and a $N \times K$ matrix $G_i(U, W)$ representing users accessing WAs.

A. CELLULAR PATHS SET DISCOVERY

A cellular path requires only one user's participation. This user is responsible for both data acquiring and uploading. We assume cellular networks fully covered the sensing area. The users encountering the PoIs in the data collecting cycle can get the data of PoIs and upload data to the back-end platform through cellular networks immediately. We assume the probability of successful data transmission through cellular networks is 100% ($q_c = 1$) without losing generality.

The candidate cellular paths set is constructed through an iterative process from time interval 1 to time interval T_n in each data collecting cycle. The basic idea is to find user-PoI encounters, add these found users to the paths set, and calculate the corresponding data delivery probabilities.

In each iteration, the following steps are repeated for each PoI l_j.

1) Scan $G_i(U, L)$ to find the users encountering the PoI l_j;
2) If it is the first time for user k encountering PoI l_j, add a new path $\{l_j: u_k: bss_e\}$ into the candidate paths set P and calculate the data delivery probability of this path as $q_{pl} = g_i(u_k, l_j)$;
3) Otherwise, P already contains this path, updates the data delivery probability of this path as

$$q_{pl} = 1 - (1-q_{pl})(1 - g_i(u_k, l_j)). \tag{6}$$

New user-PoI encountering improves the data delivery probability of the existing path. $(1 - q_{pl})(1 - g_i(u_k, l_j))$ represents the possibility that data cannot be uploaded to the back-end system through this path with new adding encountering. So the new delivery probability is $1 - (1-q_{pl})(1 - g_i(u_k, l_j))$.

Suppose the number of users is N, and the number of PoIs is L. This process needs to traverse L PoI entries per user per time interval, which leads to the time complexity of $O(T_nNL)$.

B. OPPORTUNISTIC PATHS SET DISCOVERY

An opportunistic path usually involves more than one user. The user encountering a PoI can get its data. Then through users' movement and encounter, the data can be forwarded to a WA and uploaded to the back-end platform. The participating users take the different roles of acquiring, relaying, and uploading data. We assume data can be transmitted successfully from one user to the other user/WA during their encounters without losing generality. We repeat the following steps upon each time interval for each PoI to find candidate opportunistic paths.

1) Search for possible encounters between users and PoIs by scanning $G_i(U, L)$. A none-zero value of $g_i(u_k, l_j)$ implies that the data of PoI l_j can be obtained from user u_k at the time interval i. There are two possibilities:

a) It is the first time for user k encountering PoI l_j, then add a new path $\{l_j: u_k\}$ into the candidate paths set P and calculate the data delivery probability of this path as $q_{pl} = g_i(u_k, l_j)$. This kind of path is called a partial path since the connection to a WA has not been established.
b) There is a partial path $\{l_j: u_k\}$ in the P due to an encounter between u_k and l_j in the past. In this case, the data delivery probability of this partial path is updated as $1 - (1-q_{pl})(1 - g_i(u_k, l_j))$.

2) Search for user-user encounters giving rise to new possible paths by scanning $G_i(U, U)$. A none-zero value of $g_i(u_k, u_o)$ implies that the data can be transferred from user u_k to user u_o at the time interval i. For every partial path p already in the P, if we can find a user u_o encountering the last user of this path and this user is not included in this path, a new partial path $\{l_j, u_o\}$ is inserted into P. The data delivery probability is calculated as

$$q_{pl} = q_{pl} g_i(u_k, u_o)). \tag{7}$$

There is also the possibility that the last two users of an existing partial path encounter again. It will increase the data delivery probability of this path. For example, a partial path $\{l_j: u_k, u_o\}$ was inserted into P in the past time intervals. Now users u_k and u_o encounter again at the current time interval. In this case, $g_{past}(u_k, u_o)$, the encountering probability between the ultimate and the penultimate nodes over the past time intervals that has already been factored in the computation of the data delivery probability should be removed, which leads to q_{pl} being updated to $q_{pl} = q_{pl}/g_{past}(u_k, u_o)$. Then the data delivery probability of this partial path is updated by a new value inflated by the probability of an encounter over the current time interval, $1 - (1 - g_{past}(u_k, u_o))(1 - g_i(u_k, u_o))$. Therefore, the new q_{pl} is computed as

$$q_{pl} = \frac{q_{pl}(1 - (1 - g_{past}\,(u_k, u_o)\,(1 - g_i\,(u_k, u_o)))}{g_{past}\,(u_k, u_o)}. \tag{8}$$

3) Search for possible encounters between users and WAs by scanning $G_i(U, W)$. A none-zero value of $g_i(u_o, w_m)$ implies that the data can be transferred from user u_o to WA w_m at the time interval i, which means data can be uploaded to the back-end platform. We call this kind of path as a full path. There are two possibilities.

a) For a partial path p already in the P, if the last user u_o of this path encounters a WA w_m at this time interval, a new full path $\{l_j: u_k, \ldots, u_o: w_m\}$ is inserted into P. The data delivery probability is calculated as

$$q_{pl} = q_{pl} g_i(u_o, w_m)). \tag{9}$$

b) For a full path p already in the P, if the last user u_o of this path re-encounters the same WA w_m again at this time interval, the data delivery probability of this full path should cumulate the probability of this encounter as

$$q_{pl} = \frac{q_{pl}(1 - (1 - g_{past}\,(u_o, w_m)\,(1 - g_i\,(u_o, w_m)))}{g_{past}\,(u_o, w_m)}. \tag{10}$$

Suppose the number of users is N, the number of PoIs is L, and the number of WAs is K, the first task searching for user-PoI encounters requires traversing L PoI entries per user per time interval, which leads to time complexity of $O(T_nNL)$. The second task searching for user-user encounters involves traversing N users per existing partial paths per time interval. In the worst case, traversing existing partial paths is $O(LN^{Tn-2})$. The overall time complexity of the second task is $O(LN^{Tn-1})$. The last task requires traversing K WAs per existing partial paths per time interval, which leads to the worst time complexity $O(LKN^{Tn-2})$. The time complexity of the second and last tasks can be reduced when the hop count of opportunistic paths is bounded. Even when the opportunistic protocol does not set a hardbound on the hop

count of paths, we can choose to filter out paths with a hop count higher than some threshold. We also can filter paths with data delivery probability lower than some threshold. The above filtering operations can reduce the time complexity significantly.

Furthermore, users can be divided into serval groups based on their historical traces. The users with no or little possibilities to encounter are put in different groups. When searching for possible encounters, only the same groups' users are traversed, which can also reduce the time complexity.

V. HEURISTIC ALGORITHM

Optimal data collection can be considered as choosing specific data collecting paths from candidate paths set to minimize the total crowdsensing cost under the constraints of sensing PoIs and transferring sensing data to the back-end platform. If such a set of data collecting paths is determined, all the users contained in these paths are selected to participate in crowdsensing activities. We format this problem as the following.

$$min \sum\nolimits_{p\in P} y_p c_p \tag{11}$$

$$s.t.\ y_p \in \{0,1\}, \quad \forall p \in P \tag{12}$$

$$D_{l,DCP_l} \geq D_{thre}, \quad \forall l \in L \tag{13}$$

$P = \cup_{l\in L} DCP_l$ is the candidate set of data collecting paths constructed by analyzing the historic users' traces, as discussed in the previous section. Each element $p \in P$ is a data collecting path that can get a PoI's data and send it to the back-end platform. c_p is the cost of corresponding path p, which can be calculated by Formulation (1) or (3) based on its type. y_p is a binary variable whose value only can be 0 or 1 as illustrated in the first constraint (Formulation 11). y_p decides whether the corresponding path p is selected. If this path is selected, $y_p = 1$, otherwise $y_p = 0$.

The second constraint (Formulation 12) ensures every PoI should be covered, and its data can be obtained and sent to the back-end platform with a probability higher than a threshold value. $D_{l,DCPl}$ is the probability, and D_{thre} is the threshold value.

If there is only one path p covering PoI l, the data delivery probability of PoI l is

$$D_{l,DCP_l} = q_{pl}. \tag{14}$$

If two or more paths cover PoI l, $D_{l,DCPl}$ is determined by the following steps.

1) If these paths are independent (containing no common users),
$$D_{l,DCP_l} = 1 - \prod\nolimits_{p\in DCP_l} (1 - q_{pl}). \tag{15}$$
2) If these paths are not independent (containing common users), path similarity s_p is introduced to calculate the final delivery probability.
$$s_p = \frac{\text{the number of common users}}{\text{the maximum number of users in one path}} \tag{16}$$

Heuristic Algorithm for Paths Selection

Input: candidate paths set P, PoIs set L, delivery probability threshold D_{thre}
Output: selected paths set Q, total cost $Cost(P)$

1: $Cost\ (P) = 0$; all y_p are set to 0,
2: $Q \leftarrow \emptyset$; $D_{l,DCP_l} = 0, \forall l \in L$;
3: *while* $\exists l \in L$: $D_{l,DCP_l} < D_{thre}$ *do*:
4: $\quad p \leftarrow \arg\min_{P\in P\backslash Q} \left[c_p/q_{pl}\right]$;
5: $\quad Q \leftarrow Q \cup \{p\}$;
6: $\quad$ update D_{l,DCP_l};
7: *while* $\exists p \in Q$ *do* $y_p = 1$;
8: $Cost\ (P) = \sum_{p\in P} y_p c_p$;
9: *return*

$$D_{max} = 1 - \prod\nolimits_{p\in DCP_l} (1 - q_{pl}) \tag{17}$$

$$D_{min} = max_{p\in DCP_l} q_{pl} \tag{18}$$

$$D_{l,DCP_l} = D_{min} + (D_{max} - D_{min})(1 - s_p) \tag{19}$$

This problem is a minimum set covering problem, involving linear constraints along with cost function. Since the minimum set covering problem [24] is a well-known NP-hard problem, an approximate algorithm is needed to tackle it.

The objective function, $C = min \sum_{p\in P} y_p c_p$, is a submodular function over the space of feasible solutions. In particular, for any two subsets Q_1, Q_2 of P, C satisfies $C(Q_1 \cup Q_2) + C(Q_1 \cap Q_2) = C(Q_1,) + C(Q_2)$. For the generic set cover with a submodular cost function, the recent primal-dual algorithm [43] yields a Δ approximation, where Δ corresponds to the maximum number of variables in each linear covering constraint. In our problem, the number of variables per constraint equation equals the number of DCPs per PoI. This is highly variable and grows fast with the number of mobile users and the hop count of the respective DCPs.

Therefore, we propose a greedy heuristic algorithm that constructs the solution in multiple steps by making a locally optimal decision in each step. The locally optimal decision means the desirability of including a path to cover PoI increases with its cost-effectiveness (the ratio of delivery probability and cost). The pseudo-code of the proposed algorithm is as follows:

The variables are initialized in the first and second lines. Selected paths set Q is set to empty. For each PoI j, corresponding D_{l,DCP_l} is set to 0. The lines from 3 to 6 describe the iterative procedure finding near-optimal paths set. For each PoI l, the algorithm selects the path p with the minimum ratio c_p/q_{pl} from unselected paths set $P\backslash Q$ first, adds this path to Q, and removes it from P as shown in lines 4 and 5. Then data delivery probability D_{l,DCP_l} is updated according to the process described by Formulation 13-18. If D_{l,DCP_l} is higher than the threshold value, the iterative process for this PoI ends. The algorithm always selects the path minimizing the cost per delivery probability over the set of PoIs it covers with respect to already selected paths. In lines 7 and 8, y_p is determined, and $Cost(P)$ is calculated based on Q's composition.

The algorithm is a straightforward adaptation of the well-known greedy Set Covering heuristic by Chvátal [24]. The approximation ratio is independent of the number of data collecting paths. This renders our algorithm more robust than the primal-dual one of [43] in terms of worst-case performance.

For each step s to cover PoIs, our algorithm always chooses the path with the smallest cost-effectiveness. Since the cost function is submodular, it increases the current (partial) solution's cost by at most $r_s * OPT$, where OPT denotes the optimum solution and

$$r_s = \frac{\text{the total coverage needed-the coverage already achieved}}{\text{the total coverage needed}}. \tag{20}$$

The total cost can be calculated as the following.

$$C = \sum_{s=1}^{S} C_s \leq OPT * \sum_{s=1}^{S} r_s \tag{21}$$

The approximation ratio is $\sum_{s=1}^{S} r_s$. Here, S is the number of steps needed to achieve the total coverage. Its value can be approximately calculated as

$$S \approx L * \left\lceil D_{thre}/min(q_{pl}, p \in P) \right\rceil, \tag{22}$$

where L is the number of PoIs, D_{thre} is the probabilistic delivery threshold, $min(q_{pl}, p \in P)$ is the minimum delivery probability over all the candidate data collecting paths, and $\lceil\rceil$ is the ceiling operation returning the smallest integral value that is greater than or equal to its input value.

Since r_s is less than or equal to 1, the approximation ratio is less than or equal to $L * \left\lceil D_{thre}/min(q_{pl}, p \in P) \right\rceil$ in the worst case. In the average case, $\sum_{s=1}^{S} r_s$ can be roughly considered as $\ln(L * \left\lceil D_{thre}/min(q_{pl}, p \in P) \right\rceil)$.

A quick sorting algorithm is used to sort candidate paths by ascending order of cost per delivery probability. The time complexity of the proposed heuristic algorithm is $O(LN_{cp}^2 logN_{cp})$, where L is the number of PoIs, and N_{cp} is the size of candidate paths set.

VI. PERFORMANCE EVALUATION

In this section, we carry out real-world trace-driven simulations to evaluate the proposed data collection method's performance. The results are given and discussed.

A. SIMULATION SETTINGS

We use three experimental traces, referred to as Cambridge, Infocom06, and UPB, to emulate the way nodes encounter with each other and hit the PoIs and WAs. These traces record the contact history of users carrying mobile devices. The devices periodically detect their neighbors through D2D networking interfaces and record the contact information, including two contact parties, the start time, and the duration. Cambridge and Infocom06 traces are Bluetooth based contact traces collected by the Haggle Project [25], and UPB trace is WiFi based contact trace.

The Infocom06 trace was collected over an interval of 4 days during Infocom 2006 in Barcelona. It involves 78 mobile iMotes carried by the students and researchers participating in the conference and the 20 stationary iMotes deployed at various places in the conference hotel such as conference rooms, the bar, the concierge, and the hotel elevators. The mobile iMotes have a wireless range of around 30 meters. The stationary iMotes have a more powerful battery and extended radio range (around 100 meters).

The Cambridge trace was collected through an experiment conducted for approximately 2 months in the city of Cambridge. Mobile users in this experiment consisted of 36 students from Cambridge University who were asked to carry the iMotes with them at all times for the experiment's duration. In addition to this, 18 stationary iMotes were deployed in various locations that many participants were expected to visit such as computer lab, grocery stores, pubs, market places, and shopping centers in and around Cambridge, UK. The contacts between different mobile users, and also contacts between mobile users and various fixed locations were recorded.

For evaluation purposes, the mobile iMotes are mapped to mobile users, and the stationary iMotes are mapped to PoIs. Since free WiFi access is often provided around particular locations, 50% of stationary nodes are also mapped to WAs. We assume the cellular networks fully cover the experimental area.

UPB trace was collected through an experiment that lasted 63 days at the University Politehnica of Bucharest. It involves 72 participants being students from the facility, as well as teachers and assistants, out of which only 42 had at least one contact. These 42 participants are mapped to mobile users. Based on social interaction analysis, 5 PoIs and 5 WAs are added.

We vary the data collecting cycle by extracting and working with varying-length parts of the traces. In the Infocom2006 trace, people all attend the same event, and thus they tend to fall into the same community. In the UPB trace, people are in the same facility and form several relatively stable communities. Compared to the Cambridge trace, the Infocom2006 and UPB trace have much higher network density and contact rate. Therefore, the data collecting cycle is in the order of hours for the denser Infocom2006 and UPB trace, and the data collecting cycle is in the order of days for the sparser Cambridge trace. More specifically, we change the data collecting cycle from 0.5 to 3 hours Infocom06 and UPB trace, and from 12 to 72 hours in Cambridge trace. If the data collecting cycle is too short, opportunistic paths are rare, and most devices will upload the data through cellular networks. Our setting is similar to the setting used in the comparison method [17].

The number of time intervals in a data collecting cycle is set to 5. So the maximum hops of an opportunistic path is 4. We also change the number of time intervals from 2 to 7 to study its impact on the performance.

Based on the existing research on the energy consumption in mobile devices [18], [31], [32], the energy consumption for sampling data from sensors like accelerometer and digital pressure/temperature sensor is negligible concerning the

TABLE 2. Simulation parameters.

Parameter	Traces		
	Infocom06	Cambridge	UPB
D2D interfaces	Bluetooth	Bluetooth	WiFi
# of users	78	36	42
# of WAs	10	9	5
# of PoIs	10	9	5
Trace duration(days)	4	60	63
Data collecting cycle	0.5~3	12~72	0.5~3
# of time intervals	2~7	2~7	2~7
Data sampling cost	0.2		
Opportunistic cost	1		
Cellular cost	10		

energy spent for communications. For sensors like GPS, microphone, and camera, the energy consumption is less than cellular/WiFi networks. Therefore, the cost of sampling data is significantly lower than the cost of transferring data. According to the above research, for data communications, cellular networks consume 2-4 times more energy than WiFi networks. Bluetooth consumes much less energy compared with WiFi/cellular networks. Cellular networks always cause a fee no matter what the price plan is. Therefore, we set the cost of sampling data, transferring data through one opportunistic hop, and cellular network to 0.2, 1, and 10, respectively.

The detailed simulation parameters are listed in Table 2.

B. EVALUATION METRICS

To evaluate the performance of the data collection scheme, we use the following two metrics: 1) Delivery probability: the average probability of successfully sampling and uploading PoIs' data to the back-end platform in the data collecting cycle. It indicates PoI coverage. 2) Delivery cost: the sum of all selected path costs.

C. SCHEMES FOR COMPARISON

To better understand the performance of our scheme, we compare our scheme to the three existing schemes, which aim to minimize the overall data collection cost with guaranteed PoIs data delivery.

The first scheme assumes all the participants using cellular networks to upload data to the back-end platform, referred to as CMinCost. It transfers and formats data collection problem as a minimum cost set cover problem with a submodular objective function and adopts a simple iterative process based on greedy algorithms to get the approximate optimal solution. The basic procedure is similar to CrowdRecruiter [8], [9], except piggyback mechanism is not used here.

The second scheme [17] uses an opportunistic network to upload data to the back-end platform, referred to as OMinCost. It translates the statistics of individual user mobility to statistics of space-time path formation and selects the data collection paths set with the minimum cost to meet PoIs data delivery constraints.

The third scheme [18] works in a distributed fashion and aims to minimize the cost of sensing and uploading for

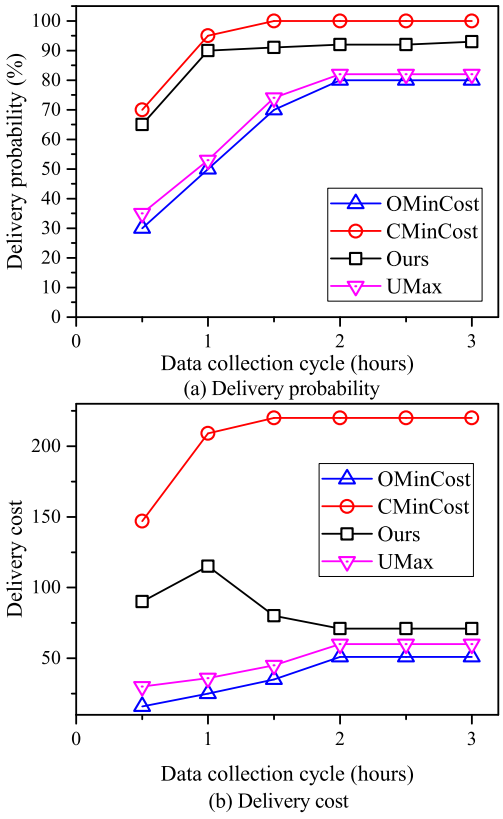

FIGURE 2. Performance comparisons on the Infocom06 trace set.

the participants, while maximizing data collection utility, referred to as UMax. Each mobile user computes a utility value based on its resource consumption (cost), location, and mobility pattern. Sensing and uploading operations occur when the utility value exceeds a threshold.

D. EVALUATION RESULTS

1) PERFORMANCE COMPARISON RESULTS

Fig. 2, Fig. 3, and Fig.4 show the performance comparing results on Infocom2006 trace, Cambridge trace, and UPB trace, respectively.

As shown in Fig.2 (a), Fig.3 (a), and Fig.4(a), CMinCost scheme always outperforms OMinCost scheme, UMax scheme, and our scheme in delivery probability. The reason is that CMinCost transfers data to the back-end platform through cellular networks. If a mobile user can visit a PoI in a data collecting cycle, the data of this PoI can be successfully uploaded to the back-end platform.

For OMinCost scheme, besides mobile users visiting PoI, it still needs to find an opportunistic path between the visiting user and WA to transfer the data to the back-end platform. There are two cases. One is failing to find an opportunistic path, which causes data not to be transferred to the back-end platform and leads to a significant low data delivery probability. The other is successfully finding an opportunistic path and transferring data to the back-end platform through this path. However, the successful data transferring probability of the opportunistic path is lower than that of the cellular path, which also hurts the final data delivery probability.

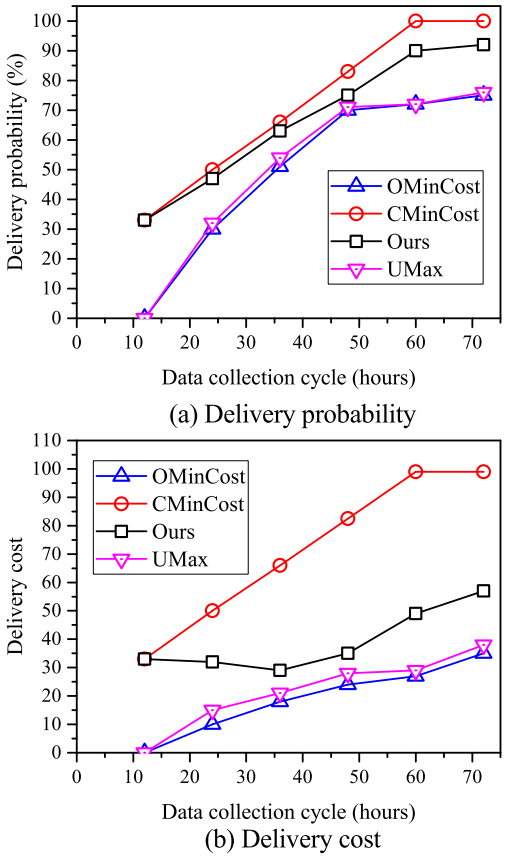

FIGURE 3. Performance comparisons on the Cambridge trace set.

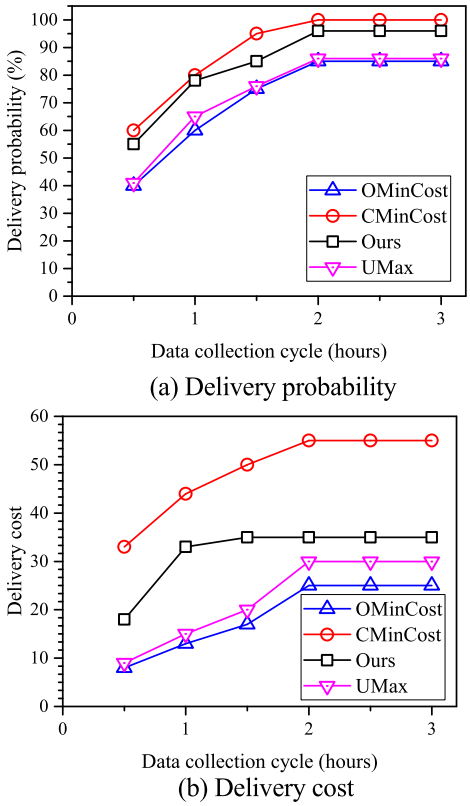

FIGURE 4. Performance comparisons on the UPB trace set.

Therefore, OMinCost always has the lowest data delivery probability in these four schemes.

Similar to OMinCost, UMax utilizes opportunistic networks to upload data. Unlike OMinCost discovering all possible opportunistic paths and selecting optimized paths from them, UMax works in a distributed fashion, which means each user decides for himself whether to join sensing and transferring task base on a utility composite metric. Compare with OMinCost, UMax usually recruits more users, which leads to a slightly higher delivery probability.

The delivery probability of our scheme is close to CMinCost since our scheme uses the cellular networks to upload data when an opportunistic path is not available. A performance margin exists between our scheme and CMinCost due to the reason that our scheme prefers to use the low-cost opportunistic path with lower successfully transferring probability compared with the cellular path.

From Fig.2 (b), Fig.3 (b), and Fig.4 (b), we also find low-cost opportunistic paths significantly reduce the delivery cost. CMinCost always has the highest cost, and OMinCost always has the lowest cost. The cost of UMax is slightly higher than OMinCost since it recruits more users participating in data sensing and transferring.

Compared with them, our scheme achieves a better tradeoff between delivery probability and cost. In the Infocom06 trace set, our scheme reduces the delivery cost by 67% with only an 8% drop in the delivery probability compared with CMinCost when the data collecting cycle is 2 hours. In the same circumstance, our scheme improves the delivery probability by 15% with 20 extra delivery cost units compared with OMinCost. In Cambridge trace set, our scheme reduces the delivery cost by 56% with only a 5% drop in the delivery probability compared with CMinCost when the data collecting cycle is 36 hours. In the same circumstance, our scheme improves the delivery probability by 24% with 11 extra delivery cost units compared with OMinCost.

Our scheme also adapts a wide range of application scenarios and needs. If the applications demand high delivery probability, opportunistic networking based methods cannot work. For example, in Infocom06 trace, the highest data delivery probability of opportunistic networking based methods is 50% when the data collecting cycle is 1 hour. Many applications may not work or perform poor with 50% sensing data missing. The delivery probability of cellular networking based methods can reach 70%. The delivery probability of our method can reach 66% at a significantly low cost. In Cambridge trace set, opportunistic networking based methods cannot get any data when the data collecting cycle is less than 12 hours. By integrating cellular networks and opportunistic networks, our method can work under different circumstances and application demands.

As illustrated in Fig.2, Fig.3, and Fig.4, our scheme improves the delivery probability more significantly in the Cambridge trace set. The reason is that the Cambridge trace set represents the sparser network environment where user encounters are fewer and opportunistic paths are harder

to formulate. Users in the UPB trace set are less concentrated compared with the Infocom06 trace set. The communication range of WiFi devices used in the UPB trace set is quite a bit larger than that of Bluetooth devices used in the Infocom06 trace set. These lead to similar opportunistic networking density. The trend of delivery probability and cost is similar in the UPB and Infocom06 Trace sets.

2) EFFECTS OF DATA COLLECTING CYCLE

We study the impacts of the data collecting cycle on the delivery probability and cost. Longer data collecting cycles enable the realization of more data-collection paths. As can be seen from Fig.2 (a), Fig.3 (a), and Fig.4 (a), the delivery probabilities of all four schemes increase with the data collecting cycle increasing.

When the data collecting cycle is small, the delivery probability of our scheme is close to CMinCost and outperforms OMinCost and UMax a lot. It is because most available data collecting paths are cellular paths. CMinCost and our scheme can use these cellular paths to sample and upload some PoIs' data, and OMinCost and UMax cannot use these cellular paths. Compared with OMinCost, UMax works in a distributed fashion and usually recruits more users. Therefore, the delivery cost of OMinCost is the lowest among these four schemes, as illustrated in Fig.2 (b). Fig. 3(b), and Fig.4 (b). Our scheme achieves lower cost than CMinCost since our scheme may find and use opportunistic paths instead of cellular paths to sample and upload certain PoIs' data.

When the data collecting cycle is large, OMinCost, UMax, and our scheme can find opportunistic paths for most PoIs, which leads to a significantly lower delivery cost compared with CMinCost. However, the successful data transferring probability of opportunistic path is lower than the cellular path, which causes the delivery probability of OMinCost, UMax, and our scheme is a little lower than CMinCost.

When the data collecting cycle increases, the delivery cost of CMinCost, UMax, and OMinCost increases because they use more cellular and opportunistic paths respectively to cover more PoIs. The delivery cost of our scheme may decrease at a certain point when more opportunistic paths and fewer cellular paths are used. For example, in the Infocom06 trace set, the data collecting cycle changes from 1 to 2 hours.

3) EFFECTS OF THE NUMBER OF TIME INTERVALS

Now we study the impacts of the number of time intervals in a data collecting cycle, T_n, on the performance of OMinCost and our scheme. Since UMax does not discover all possible opportunistic paths and select optimized paths from them, T_n has no impact on its performance.

As discussed in the previous sections, we divide the data collecting cycle into T_n time intervals. The maximum hop counts of an opportunistic path is $T_n - 1$, which has a great influence on the time complexity of constructing candidate paths set. It also has a direct impact on N_{cp}, the size of the candidate paths set. Generally speaking, the larger T_n is, the larger N_{cp} is. N_{cp} affects the time complexity of our heuristic algorithm.

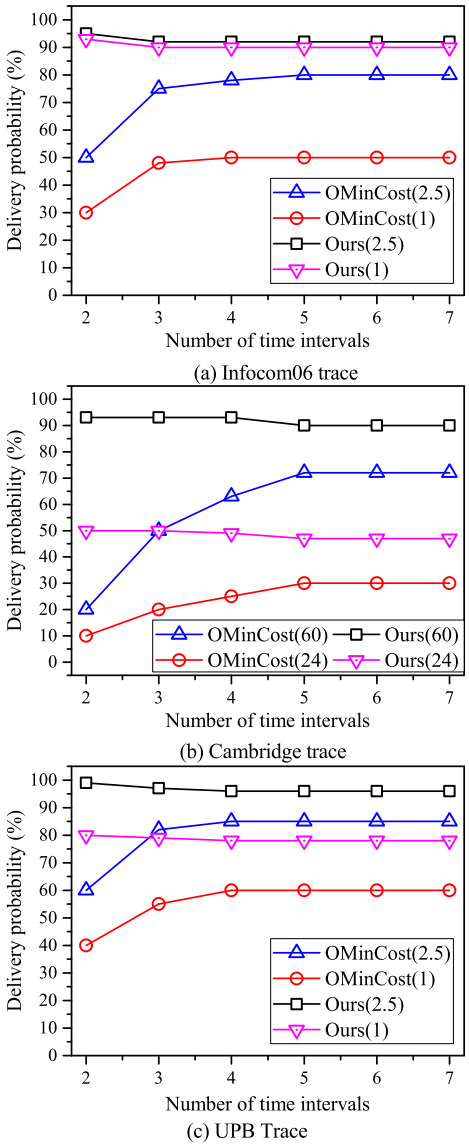

FIGURE 5. The impact of T_n on delivery probability.

With the increasing of T_n, more opportunistic paths can be found, making OMinCost achieve higher data delivery probability as shown in Fig.5. However, when T_n exceeds a particular value, the growth in data delivery probability is over. In the Infocom06 trace set, this particular value is 4 when the data collecting cycle is 1 hour or 5 when the data collecting cycle is 2.5 hours. In the Cambridge trace set, this particular value is 5. In the UPB trace set, this particular value is 4. It indicates the most useful opportunistic paths are 1, 2, or 3 hops, and 2-hop paths make up a significant part of this paths group.

When the data collecting cycle is large (2.5 hours in infocom06/UPB trace set and 60 hours in Cambridge trace set), the improvement in data delivery probability is more significant. It is because the larger data collecting cycle allows more relatively longer opportunistic paths to emerge.

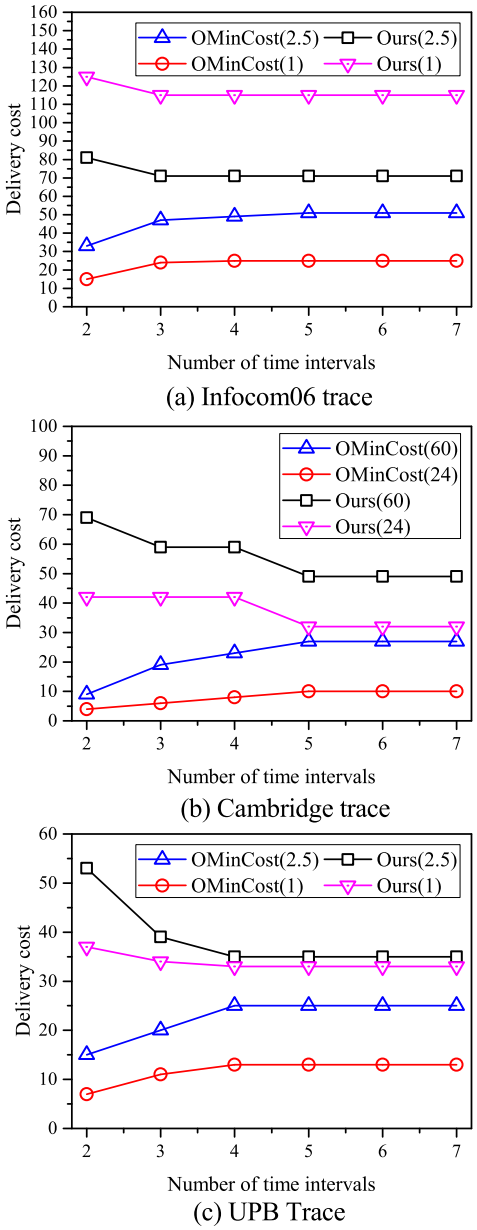

FIGURE 6. The impact of T_n on delivery cost.

For our scheme, the data delivery probability decreases slightly and then stabilizes with the increase of T_n. When T_n is small, the opportunistic paths are limited by allowed hops. For example, when T_n is 2, only 1-hop opportunistic paths can be used. Our scheme may use more reliable cellular paths to upload data, which leads to higher data delivery probability. When T_n exceeds the particular value discussed before, most opportunistic paths are available for being chosen by our scheme, which leads to a little lower delivery probability.

As shown in Fig.6, the delivery cost of OMinCost shows the same trend with delivery probability due to the same reason discussed above. It increases first and then stabilizes after T_n passes a particular value. The delivery cost of our scheme also shows the same trend with the delivery probability. It decreases first and then stabilizes after T_n passes a particular value.

Based on our experiments, most of the selected opportunistic paths are 1-hops, 2-hops, and 3-hops. A few 4-hops opportunistic paths can be used when the data collecting cycle is long. Therefore, we can set T_n to a small value, which significantly reduces the time complexity of constructing candidate paths set and make our scheme more practical.

VII. CONCLUSION

This paper studies the data collection problem for mobile crowdsensing over integrated cellular and opportunistic networks. Specially, we define optimal data collection as choosing specific data collecting paths from the candidate path set to minimize the total cost under the data delivery constraints. First, we prove such a problem is an NP-hard minimum set coverage problem. Then, the heuristic algorithm is proposed to get an approximate optimal solution. Finally, we evaluate the proposed scheme's performance through simulations on three real-world traces: Cambridge, Infocom06, and UPB. Compared with cellular networking-based approaches, our scheme reduces the cost significantly with a slight delivery loss. Compared with opportunistic networking-based approaches, our scheme significantly improves the delivery probability with a moderate cost increase. By integrating the advantages of cellular and opportunistic networks, our scheme can work under different circumstances and application demands and provide a better tradeoff between delivery probability and cost.

Now we assume all the devices have the same cellular transferring cost. Due to the vast existence of heterogeneous cellular cost model, we plan to consider heterogeneous cost model and hybrid data collecting paths in the future work. We also plan to incorporate learning assisted users' movement predicting, cooperative caching, and data aggregating mechanisms to improve performance.

REFERENCES

[1] J. Liu, H. Shen, H. S. Narman, W. Chung, and Z. Lin, "A survey of mobile crowdsensing techniques: A critical component for the Internet of Things," *ACM Trans. Cyber-Phys. Syst.*, vol. 2, no. 3, pp. 1–26, Jul. 2018, doi: 10.1145/3185504.

[2] J. Wang, L. Wang, Y. Wang, D. Zhang, and L. Kong, "Task allocation in mobile crowd sensing: State-of-the-art and future opportunities," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3747–3757, Oct. 2018, doi: 10.1109/JIOT.2018.2864341.

[3] J. Dutta, C. Chowdhury, S. Roy, A. I. Middya, and F. Gazi, "Towards smart city: Sensing air quality in city based on opportunistic crowd-sensing," in *Proc. 18th Int. Conf. Distrib. Comput. Netw. (ICDCN)*, 2017, pp. 1–6, doi: 10.1145/3007748.3018286.

[4] Y. Xu, Y. Zhu, and Z. Qin, "Urban noise mapping with a crowd sensing system," *Wireless Netw.*, vol. 25, no. 5, pp. 2351–2364, Jul. 2019, doi: 10.1007/s11276-018-1663-x.

[5] V. Freschi, S. Delpriori, L. C. Klopfenstein, E. Lattanzi, G. Luchetti, and A. Bogliolo, "Geospatial data aggregation and reduction in vehicular sensing applications: The case of road surface monitoring," in *Proc. Int. Conf. Connected Vehicles Expo (ICCVE)*, Vienna, Austria, Nov. 2014, pp. 711–716, doi: 10.1109/ICCVE.2014.7297643.

[6] Z. Peng, X. Gui, J. An, T. Wu, and R. Gui, "Multi-task oriented data diffusion and transmission paradigm in crowdsensing based on city public traffic," *Comput. Netw.*, vol. 156, pp. 41–51, Jun. 2019, doi: 10.1016/j.comnet.2019.03.020.

[7] X. Wang, W. Wu, and D. Qi, "Mobility-aware participant recruitment for vehicle-based mobile crowdsensing," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4415–4426, May 2018, doi: 10.1109/TVT.2017.2787750.

[8] H. Xiong, D. Zhang, Z. Guo, G. Chen, and L. E. Barnes, "Near-optimal incentive allocation for piggyback crowdsensing," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 120–125, Jun. 2017, doi: 10.1109/MCOM.2017.1600748.

[9] D. Zhang, H. Xiong, L. Wang, and G. Chen, "CrowdRecruiter: Selecting participants for piggyback crowdsensing under probabilistic coverage constraint," in *Proc. ACM Int. Joint Conf. Pervas. Ubiquitous Comput. UbiComp Adjunct*, Seattle, WA, USA, 2014, pp. 703–714, doi: 10.1145/2632048.2632059.

[10] P. Huang, W. Zhu, K. Liao, T. Sellis, Z. Yu, and L. Guo, "Efficient algorithms for flexible sweep coverage in crowdsensing," *IEEE Access*, vol. 6, pp. 50055–50065, 2018, doi: 10.1109/ACCESS.2018.2868931.

[11] H. Xiong, D. Zhang, G. Chen, L. Wang, and V. Gauthier, "CrowdTasker: Maximizing coverage quality in piggyback crowdsensing under budget constraint," in *Proc. IEEE Int. Conf. Pervas. Comput. Commun. (PerCom)*, St. Louis, MO, USA, Mar. 2015, pp. 55–62, doi: 10.1109/PERCOM.2015.7146509.

[12] N. D. Lane, Y. Chon, L. Zhou, Y. Zhang, F. Li, D. Kim, G. Ding, F. Zhao, and H. Cha, "Piggyback CrowdSensing (PCS): Energy efficient crowdsourcing of mobile sensor data by exploiting smartphone app opportunities," in *Proc. 11th ACM Conf. Embedded Networked Sensor Syst. (SenSys)*, Roma, Italy, 2013, pp. 1–14, doi: 10.1145/2517351.2517372.

[13] N. Chakchouk, "A survey on opportunistic routing in wireless communication networks," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2214–2241, 4th Quart., 2015, doi: 10.1109/COMST.2015.2411335.

[14] J. Hu, L.-L. Yang, K. Yang, and L. Hanzo, "Socially aware integrated centralized infrastructure and opportunistic networking: A powerful content dissemination catalyst," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 84–91, Aug. 2016, doi: 10.1109/MCOM.2016.7537181.

[15] L. Wang, D. Zhang, H. Xiong, J. P. Gibson, C. Chen, and B. Xie, "EcoSense: Minimize participants' total 3G data cost in mobile crowdsensing using opportunistic relays," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 47, no. 6, pp. 965–978, Jun. 2017, doi: 10.1109/TSMC.2016.2523902.

[16] G. S. Tuncay, G. Benincasa, and A. Helmy, "Participant recruitment and data collection framework for opportunistic sensing: A comparative analysis," in *Proc. 8th ACM MobiCom Workshop Challenged Netw. (CHANTS)*, Miami, FL, USA, 2013, p. 25, doi: 10.1145/2505494.2505502.

[17] M. Karaliopoulos, O. Telelis, and I. Koutsopoulos, "User recruitment for mobile crowdsensing over opportunistic networks," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Hong Kong, Apr. 2015, pp. 2254–2262, doi: 10.1109/INFOCOM.2015.7218612.

[18] A. Capponi, C. Fiandrino, D. Kliazovich, and P. Bouvry, "Energy efficient data collection in opportunistic mobile crowdsensing architectures for smart cities," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Atlanta, GA, USA, May 2017, pp. 307–312, doi: 10.1109/INFCOMW.2017.8116394.

[19] M. Zhang, P. Yang, C. Tian, S. Tang, X. Gao, B. Wang, and F. Xiao, "Quality-aware sensing coverage in budget-constrained mobile crowdsensing networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 7698–7707, Sep. 2016, doi: 10.1109/TVT.2015.2490679.

[20] J. Liu, L. Bic, H. Gong, and S. Zhan, "Data collection for mobile crowdsensing in the presence of selfishness," *EURASIP J. Wireless Commun. Netw.*, vol. 2016, no. 1, p. 82, Dec. 2016, doi: 10.1186/s13638-016-0580-x.

[21] P. Nguyen and K. Nahrstedt, "Context-aware crowd-sensing in opportunistic mobile social networks," in *Proc. IEEE 12th Int. Conf. Mobile Ad Hoc Sensor Syst.*, Dallas, TX, USA, Oct. 2015, pp. 477–478, doi: 10.1109/MASS.2015.80.

[22] E. Wang, Y. Yang, J. Wu, W. Liu, and X. Wang, "An efficient prediction-based user recruitment for mobile crowdsensing," *IEEE Trans. Mobile Comput.*, vol. 17, no. 1, pp. 16–28, Jan. 2018, doi: 10.1109/TMC.2017.2702613.

[23] E. Wang, Y. Yang, and K. Lou, "User recruitment for optimizing requester's profit in self-organized mobile crowdsensing," *IEEE Access*, vol. 6, pp. 17518–17526, 2018, doi: 10.1109/ACCESS.2018.2814739.

[24] V. Chvatal, "A greedy heuristic for the set-covering problem," *Math. Oper. Res.*, vol. 4, no. 3, pp. 233–235, Aug. 1979, doi: 10.1287/moor.4.3.233.

[25] (May 2009). *CRAWDAD Dataset Cambridge/Haggle (v. 2009-05-29*. [Online]. Available: httpscrawdadorgcambridgehaggle20090529imote Httpsdoiorg1015783C70011

[26] B. A. Coll-Perales, J. Gozalvez, and J. L. Maestre, "5G and beyond: Smart devices as part of the network fabric," *IEEE Netw.*, vol. 33, no. 4, pp. 170–177, Jul. 2019, doi: 10.1109/MNET.2019.1800136.

[27] Y. Liu, H. Wang, M. Peng, J. Guan, J. Xu, and Y. Wang, "DeePGA: A privacy-preserving data aggregation game in crowdsensing via deep reinforcement learning," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 4113–4127, May 2020, doi: 10.1109/JIOT.2019.2957400.

[28] Y. Liu, W. Quan, T. Wang, and Y. Wang, "Delay-constrained utility maximization for video ads push in mobile opportunistic D2D networks," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 4088–4099, Oct. 2018, doi: 10.1109/JIOT.2018.2849007.

[29] Y. Liu, L. Hao, Z. Liu, K. Sharif, Y. Wang, and S. K. Das, "Mitigating interference via power control for two-tier femtocell networks: A hierarchical game approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 7194–7198, Jul. 2019, doi: 10.1109/TVT.2019.2916715.

[30] B. Coll-Perales, L. Pescosolido, A. Passarella, J. Gozalvez, and M. Conti, "Opportunistic D2D-aided uplink communications in 5G and beyond networks," in *Wired/Wireless Internet Communications*, vol. 11618, M. Di Felice, E. Natalizio, R. Bruno, and A. Kassler, Eds. Cham, Switzerland: Springer, 2019, pp. 141–153.

[31] X. Chen, N. Ding, A. Jindal, Y. C. Hu, M. Gupta, and R. Vannithamby, "Smartphone energy drain in the wild: Analysis and implications," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 43, no. 1, pp. 151–164, Jun. 2015, doi: 10.1145/2796314.2745875.

[32] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani, "Energy consumption in mobile phones: A measurement study and implications for network applications," in *Proc. 9th ACM SIGCOMM Conf. Internet Meas. Conf. (IMC)*, Chicago, IL, USA, 2009, p. 280, doi: 10.1145/1644893.1644927.

[33] R.-C. Marin, C. Dobre, and F. Xhafa, "Exploring predictability in mobile interaction," in *Proc. 3rd Int. Conf. Emerg. Intell. Data Web Technol.*, Bucharest, Romania, Sep. 2012, pp. 133–139, doi: 10.1109/EIDWT.2012.29.

[34] (Oct. 2016). *CRAWDAD Dataset Upb/Hyccups (v. 2016-10-17)*. [Online]. Available: https://crawdad.org/upb/hyccups/20161017, doi: 10.15783/C7TG7K.

[35] W. Wang, V. Srinivasan, and M. Motani, "Adaptive contact probing mechanisms for delay tolerant applications," in *Proc. 13th Annu. ACM Int. Conf. Mobile Comput. Netw. (MobiCom)*, Montreal, QC, Canada, 2007, p. 230, doi: 10.1145/1287853.1287882.

[36] J. Zhao, X. Zhuo, Q. Li, W. Gao, and G. Cao, "Contact duration aware data replication in DTNs with licensed and unlicensed spectrum," *IEEE Trans. Mobile Comput.*, vol. 15, no. 4, pp. 803–816, Apr. 2016, doi: 10.1109/TMC.2015.2439271.

[37] S. Roy, N. Ghosh, P. Ghosh, and S. K. Das, "BioMCS: A bio-inspired collaborative data transfer framework over fog computing platforms in mobile crowdsensing," in *Proc. 21st Int. Conf. Distrib. Comput. Netw.*, Kolkata, India, Jan. 2020, pp. 1–10, doi: 10.1145/3369740.3369788.

[38] C. Piao and C. H. Liu, "Energy-efficient mobile crowdsensing by unmanned vehicles: A sequential deep reinforcement learning approach," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6312–6324, Jul. 2020, doi: 10.1109/JIOT.2019.2962545.

[39] J. Yang, L. Fu, B. Yang, and J. Xu, "Participant service quality aware data collecting mechanism with high coverage for mobile crowdsensing," *IEEE Access*, vol. 8, pp. 10628–10639, 2020, doi: 10.1109/ACCESS.2020.2965734.

[40] X. Tao and W. Song, "Task allocation for mobile crowdsensing with deep reinforcement learning," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Seoul, South Korea, May 2020, pp. 1–7, doi: 10.1109/WCNC45663.2020.9120489.

[41] W. Liu, Y. Yang, E. Wang, and J. Wu, "User recruitment for enhancing data inference accuracy in sparse mobile crowdsensing," *IEEE Internet Things J.*, vol. 7, no. 3, pp. 1802–1814, Mar. 2020, doi: 10.1109/JIOT.2019.2957399.

[42] E. Wang, Y. Yang, J. Wu, K. Lou, D. Luan, and H. Wang, "User recruitment system for efficient photo collection in mobile crowdsensing," *IEEE Trans. Human-Machine Syst.*, vol. 50, no. 1, pp. 1–12, Feb. 2020, doi: 10.1109/THMS.2019.2912509.

[43] C. Koufogiannakis and N. E. Young, "Greedy Δ—Approximation algorithm for covering with arbitrary constraints and submodular cost," *Algorithmica*, vol. 66, no. 1, pp. 113–152, May 2013, doi: 10.1007/s00453-012-9629-3.

DOAA MOHSIN MAJEED received the B.Sc. degree in computer science from the University of Kufa, Iraq, in 2011, and the M.Sc. degree in computer science and information technology from the Sam Higginbotom Institute, India, in 2013. She is currently pursuing the Ph.D. degree in computer science and technology with the Huazhong University of Science and Technology, Wuhan, China. Her current research interests include wireless sensor networks, mobile social networks, and mobile crowdsensing.

LIN ZHANG received the B.E. degree in computer science and technology from Hunan University, Hunan, Changsha, China, in 2018. She is currently pursuing the M.Phil. degree in computer software and theory with the Huazhong University of Science and Technology, Hubei, Wuhan, China. Her current research interest includes mobile crowdsensing.

KE SHI (Member, IEEE) received the B.S. and Ph.D. degrees from the Huazhong University of Science and Technology, in 1994 and 2000, respectively. From 2004 to 2006, he was with the Wireless Network Laboratory, Cornell University, as a Visiting Associate Professor. Since 2008, he has been a Full Professor with the School of Computer Science and Technology, Huazhong University of Science and Technology. His current research interests include ad hoc and wireless networks, deep learning, and big data analytics and its industrial applications.